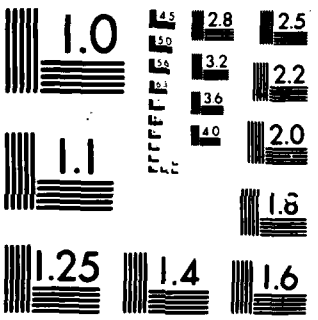
AD-A164 005 MULTI-USER REAL-TIME SPEECH PROCESSING FACILITY(U) 1/1
INDIANA UNIV AT BLOOMINGTON DEPT OF PSYCHOLOGY
D B PISONI 21 MAR 85 AFOSR-TR-86-0005 AFOSR-83-0218
UNCLASSIFIED F/G 17/2 NL

END
FILMED
1-86
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

(2)

T DOCUMENTATION PAGE

1a. RI
UNC

AD-A164 005

1b. RESTRICTIVE MARKINGS

2a. SE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 86-0005

6a. NAME OF PERFORMING ORGANIZATION
Indiana University Foundation

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research/NL

6c. ADDRESS (City, State and ZIP Code)
P. O. Box 1847
Bloomington, IN 47402

7b. ADDRESS (City, State and ZIP Code)
Building 410
Bolling AFB, DC 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL (If applicable)
NL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-83-0218

8c. ADDRESS (City, State and ZIP Code)
Building 410
Bolling AFB DC 20332-6448

10 SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2917	A4	

11. TITLE (Include Security Classification)
Multi-User Real-Time Speech Processing Facility

12. PERSONAL AUTHOR(S)
Dr. David B. Pisoni Dept. of Psychology Indiana University

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
FINAL	FROM 6/1/83 TO 8/31/84	1985 March 21	10

16. SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Speech, synthesis, analysis, perception, recognition, I/O, human factors, cognitive processes, communication sciences

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Research projects requiring the VAX 11/750 and related peripherals have all focused on speech analysis, perception and recognition. We compared perceptual confusions occurring for natural and synthetic speech syllables which showed that synthetic speech is not equivalent to "noisy" or degraded natural speech. We conducted a study that indicated that perception of synthetic speech is improved by training. Training with fluent synthetic sentences improves performance for both isolated words and sentences. Training with isolated words improves performance on isolated words but does not improve performance on fluent synthetic sentences. A large - scale series of experiments investigated the effects of noise in a talker's ears on speech production. Words produced in noise are longer, louder and higher in pitch than words produced in the quiet. The tilt of the power spectrum decreased and formant frequencies shifted in noise. Further projects in preparation include detailed acoustic-phonetic and multi-demensional analyses of speech produced under conditions of noise, stress

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

MAY 6 1985

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE NUMBER (Include Area Code)	22c OFFICE SYMBOL
Dr. John Tangney	(202) 767-5021	NL

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

86 2 11 118

CONTINUED Block number 19 Abstract

and acceleration and additional research on perceptual and cognitive constraints imposed on the listener when listening to synthetic speech is continuing as well.

AFOSR-TR- 86-0005

Multi-user Real-time Speech Processing Facility

DoD University Instrumentation Program

Final Report: AFOSR-83-0218

Granting Agency

AFOSR/NL
Building 410
Bolling AFB, DC 20332

Date: March 21, 1985

Institution

Indiana University Foundation
P. O. Box 1847
Bloomington, Indiana 47402
(812) 335-3961

Principal Investigator

David B. Pisoni, Ph.D.
Professor of Psychology
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405
(812) 335-1155

David B Pisoni

David B. Pisoni, Principal Investigator

Multi-user Real-time Speech Processing Facility:

Final Report

Equipment purchased under the DoD Instrumentation Grant

(AFOSR-83-0218)

Item:	Manufacturer:	Cost:
Mainframe		
VAX 11/750 System 2 RA-81 disks	Digital Equipment Corp.	$112,031.45
2 Mbytes Memory	National Semi-Conductor	3,221.50
4 VT100AA Terminals	Digital	8,652.09
Retrographics Upgrade Kits	Digital Engineering	5,558.32
Communications Controller Distribution Panel	Emulex	6,793.00
Expansion Box DD-11-DK	Digital Equipment	3,451.01
600 LPM Line Printer	Data Systems	9,070.00
Magnetic Tape System and Controller	Emulex & Kennedy	8,940.00
Networking Hardware DR-11-W I/O Interface WLL-11 Adapter	MDB Systems	7,624.50
VAX/VMS Documentation	Digital	661.47
Cabinets	NY Computer Exch.	1,475.00
4 Modems	Hays	1,989.00
Symbolics 3670 (partial)	Symbolics	19,004.04
Miscellaneous: shipping charges, software, adaptor kits, service charges		1,551.62
	Total:	190,023.00

Equipment Purchased with Indiana University Cost-Sharing Funds

Item:	Manufacturer:	Cost:
A/D and D/A Multi-User Audio Subsytem:		
Dual DSC-200 Analog I/O System	Digital Sound	$ 47,110.50
Interactive Laboratory System Software for Speech/Signal Processing;	Signal Technologies	12,806.50
Miscellaneous Installation Costs for VAX 11/750 2 Tape Racks,Belts,	Digital Equipment and others	1,702.00
Mag Tape Unit (partial)	Kennedy	[illegible]
REF-11 Software	DG Systems	200.00
NIL Dialect of Lisp	MIT AI Laboratory	100.00
VAX/VMS Software Update	Digital	130.35
Modem 1200	Hays	442.95
Symbolics 3670 (Partial)	Symbolics	44,[illegible]
Site Preparation Costs:		
	IU Physical Plant	1,045.34
		[illegible]
		154.45
Total:		110,[illegible]

Multi-user Real-time Speech Processing Facility:

Final Report

I. Summary of the Instrumentation Proposed and Acquired

In our original proposal for the DoD University Research Instrumentation Grant, we requested funds to purchase a specialized "state-of-the-art" signal processing facility to support a broad program of basic and applied research on human information processing and man-machine interaction using speech I/O. The proposed system was based around a VAX-11/780 computer with high-speed graphics, an array processor, and a DSC multi-channel analog-to-digital and digital-to-analog interface system.

Based on the Instrumentation Grant that we receieved, we modified the configuration of the proposed system too provide similar functionality at a reduced cost. The system that we purchased consisted of a VAX-11/750 computer with floating point accelerator, four Retrographics-fitted VT-100 terminals, and the DSC analog I/O interface system. Thus, we were able to bring up four digital speech processing stations; each station is capable of digitizing and playing speech, and can display on a raster-graphics screen speech waveforms, LPC power spectra, pseudo-spectrograms, and vocal tract cross sections. Furthermore this system can be expanded to add many more analysis workstations by simply adding more graphics terminals. In addition to the VAX, some of the DoD Instrumentation Grant funds were used to purchase a Symbolics 3670 Lisp machine to enhance our digital speech processing facilities. This system will be used to run the Spire software developed at MIT by Victor Zue and David Shipman. This software converts the Lisp machine into a powerful, highly interactive acoustic-phonetic workstation that can display high-quality spectrograms and twenty-five acoustic parameters important in the analysis of the prosodic and segmental structure of speech. Spire also links into another package called Spirex that will automatically measure and report information about the acoustic structure of segments across a large database of utterances. Spirex allows a researcher to specify an hypothesis about the acoustic-phonetic properties of speech and then test the hypotheses with data.

The VAX and the Lisp machine provide the most advanced digital speech processing facilities that are currently available. A major consideration in the choice of these two computers was the issue of compatibility with colleagues currently engaged in speech research across the country. The VAX and the Lisp machine represent two technological standards that have become significant in speech research and cognitive science and we have derived the benefit of these standards by acquiring software from a number of our colleagues to aid in speech analysis, synthesis, and cognitive simulation. Furthermore, these systems have already had a significant impact on our

research program and have made possible research that would have otherwise been inconceivable with our current PDP-11 systems alone.

Beyond the basic capabilities of speech synthesis and analysis, these systems have provided much needed new capabilities for developing and investigating large databases of speech and for modelling perceptual and cognitive processing used in spoken language understanding. Since the VAX was brought on line, it has been in constant use by undergraduate, graduate, and postdoctoral students, and by research scientists in our laboratory, as well as by colleagues at Indiana University and around the country who are interested in spoken language processing

II. DoD Sponsored Research on Speech Processing

In our original proposal, we described three projects supported by AFAMRL, Wright-Patterson AFB. The basic goal of this research is to investigate precisely why synthetic speech is so difficult to perceive and comprehend. These three projects focused on several human factors issues surrounding the use of synthetic speech in voice response systems used in cockpits. These projects all involve perceptual testing with human observers. The first project is designed to investigate differences in the perception and cognitive processing of natural and synthetic speech. The second project is directed at questions concerning the perception and encoding of synthetic speech in short- and long-term memory and the relations between processing capacity, encoding strategies, and rehearsal in STM. Finally, the third project examines a number of central issues involving the allocation of attentional resources in speech perception when the listener is required to process several simultaneous sources of information.

In the time period since we brought the VAX up into service in the lab, we have conducted a number of experiments in each of these projects. We have compared the perceptual confusions that occur for natural and synthetic syllables. When natural syllables are presented in noise, listeners confuse consonants that are phonetically similar such as /b/ and /p/. However, confusions of synthetic syllables often occur for consonants that are quite different phonetically such as /b/ and /r/. These perceptual errors do not occur with natural speech in noise. These results demonstrate that synthetic speech should not be thought of as equivalent to "noisy" or "degraded" natural speech. Rather synthetic speech is structurally very different from degraded natural speech. We have also examined the relative salience of the acoustic-phonetic structure of natural and synthetic speech using a gating procedure. In the gating procedure, listeners are presented a 50 msec segment of the waveform of a word and they are instructed to identify the word. The amount of waveform presented to listeners is increased in 50 msec steps until the word is identified. Using this technique

with natural and synthetic words has shown that natural words can be identified at much shorter durations than synthetic words. This suggests that a critical difference between natural and synthetic speech is in the richness and redundancy of the acoustic-phonetic structure -- synthetic speech is more impoverished in coarticulatory information than natural speech.

In addition to these studies, we have investigated the effects of training on perception of synthetic speech. Our results indicate that perception of synthetic speech is improved by training and experience with fluent synthetic sentences. However, we have also found that while training on isolated synthetic words will improve recognition of other isolated synthetic words, this type of training does not improve perception of fluent synthetic sentences. In contrast, training with synthetic sentences improves performance on isolated words and sentences. These results have important consequences for the design of training programs for the use of speech technology.

Other studies have examined the role of short-term memory in tasks involving synthetic speech. We have found evidence that indicates that the transfer of synthetic words from short-term memory to long-term memory is impaired relative to retention of natural speech. Most importantly, we found that this result was not specifically a function of the intelligibility of the speech but was instead a consequence of differences between natural and synthetic speech. Also, in more recent experiments, we have found that auditory storage of synthetic speech is no less stable than auditory memory for natural speech, although there are substantial differences in the encoding of these signals.

Since we acquired the VAX, we have begun a new project designed to systematically investigate the effects of noise in a talker's ears on speech production. The practical application of speech recognition as an interface to control systems in cockpits will require a better understanding of the changes that occur in speech produced in noise, under acceleration, stress and cognitive load, and in the vibration of an aircraft. It is well known that the performance of speech recognition systems degrades substantially in high-noise environments, but this has generally been attributed to the problem of detecting words in a background of noise and filtering out or extracting templates from a noise background. However, there has been very little consideration of the effects of noise on the talker who is the source of the speech signals. We have analyzed words from the Air Force speech recognition vocabulary that were produced in the quiet and in 90 dB of noise. The noise was presented over headphones to the talkers and was not recorded with the uttterances. Our analyses of speech produced in the quiet and in 90 dB of noise reveal a number of changes in the acoustic-phonetic properties of speech produced in noise. We found that words produced in noise are longer, louder, and higher in pitch than words produced in quiet; this finding replicates a number of earlier studies on the Lombard effect. However, we also found changes in the acoustic-phonetic properties of the words: the tilt of the power

spectrum decreased and formant frequencies shifted in noise. These changes indicate that speech recognition systems will not perform better in high-noise environments just by using good endpoint detection and noise-filtering algorithms; these systems must incorporate, in some fashion, knowledge of the effects of noise on the acoustic-phonetic and prosodic structure of speech.

The VAX has been an integral part of all of our work under the Air Force contract. We have used it for stimulus preparation and the acoustic analysis of speech produced in noise. As we shift more of our signal processing software from the PDP-11 systems to the VAX, it will come to play an even more important role in our research program. Projects that are underway to further investigate perceptual confusions in natural and synthetic speech rely on multidimensional analyses of responses as well as acoustic analyses of the stimuli. All these analyses will be carried out on the VAX. Studies of the attentional limitations on perception of synthetic speech and training effects and gating will make use of stimuli generated using the VAX. Finally, without a multi-user, real-time digital speech processing system, we would be unable to continue to investigate changes in speech production that result from noise, acceleration, and stress. When the Spire system is fully implemented on the Symbolics Lisp machine, it too will contribute substantially to meeting the goals of this project by allowing researchers to rapidly form and test hypotheses about changes in the acoustic-phonetic properties of speech produced under a wide variety of conditions.

III. Non-DoD Research on Speech Processing

One of the major sources of funding for the Speech Research Laboratory has come from an NIH research grant (NS-12179) to Indiana University. The research conducted under this grant covers a very wide range of projects on speech perception, analysis, and synthesis. These projects include studies of word recognition, prosody, perception of fluent, continuous speech, and simulations of the perceptual and cognitive processes that mediate spoken language understanding. The impact of the new instrumentation on all of this work has been substantial, so only a few examples will be given here. We have carried out several studies investigating the perception of natural and synthetic speech with special populations of listeners including foreign talkers of English and young children. The results of these studies have generally indicated that perception of synthetic speech is substantially impaired compared to natural speech when listeners have poor English language skills or reduced cognitive capabilities. By using the ILS digital signal analysis system on the VAX, we will be able to manipulate the structural properties of synthetic speech signals to determine how these properties interact with linguistic and cognitive capabilities.

Also, we have investigated the role of prosody in perception of fluent speech. Our initial studies have indicated that pitch inflection significantly aids word recognition in fluent synthetic sentences. We intend to extend this investigation to natural sentences as well. The VAX provides researchers with the capability of conducting LPC analyses of spoken sentences, modifying the prosodic characteristics of those sentences and then resynthesizing them as stimuli. Moreover, the multi-user facilities permit several researchers to work on these projects simultaneously.

One major research project has been concerned with the structural properties of words in the mental lexicon. As part of this project, we have been acquiring lexical databases from a number of different sources. These databases include a 250,000 word dictionary containing spellings, phonetic transcriptions, and syntactic markings, a 20,000 word subset with orthographic, phonetic, and syntactic codes, and including segmental durations and pitch information, and database of word frequency counts derived from one million words of text. In addition, we have developed several databases of children's lexicons. Together, this databases have been used to systematically investigate structural properties of words. This research has focused on issues that include optimal search strategies for recognizing words in a large vocabulary, structural accounts of the word frequency effect, and the structural characteristics of the mental lexicon as it develops with linguistic experience. These studies are all carried out on the VAX using special software developed to search and classify words in these lexical databases. Because of the immense size of these databases, this work would have been literally impossible using our PDP-11 computers. Furthermore, we have recently carried out a large-scale study to investigate the size of the mental lexicon. In this experiment we collected familiarity ratings from 600 Indiana University undergraduates for each of the words in the 20,000 word database. These ratings will provide accurate information about the lexical knowledge listeners have for a very large set of words. The ability to analyze the data using the high computational speed of the VAX and the large amount of on-line storage made this project feasible.

We have also used the VAX extensively to analyze the effects of alchohol ingestion on speech production. We analyzed a large number of words and sentences spoken by several male talkers when sober and when intoxicated. This study was designed to determine whether reliable acoustic-phonetic differences can be observed between normal and intoxicated speech conditions. At this stage in the analysis, it appears that there are large and reliable differences in the durations of segments produced by sober and intoxicated talkers. These analyses were carried out using the same digital speech processing techniques used to investigate the effects of high-levels of noise on speech production.

Recently, we have begun to develop a set of large databases of natural and synthetic speech. These stimuli are stored as digital waveform and analysis files on the VAX. One database consists of isolated MRT and PB words and spoken letters and digits produced by 10 male and female talkers. We have also begun to collect a large number of sentences and long passages of speech for use in experiments. As part of a project designed to assess the performance of commercially available text-to-speech systems, we have acquired the Texas Instruments database of 26 tokens of 46 words produced by eight male and eight female talkers. This database will also be used to develop new speech recognition algorithms based on new knowledge about human speech perception.

We will also be developing a special database of fluent speech for analysis on the Symbolics Lisp machine. The Spire system running on the Lisp machine allows researchers to transcibe speech using orthography and phonetic notation. These transcriptions are linked to the waveform and analysis parameters for each utterance. Thus, this system allows researchers to formulate and test hypotheses about the relationship between acoustic properties and linguistic context using the Spirex database system with Spire. These hypotheses can then be translated into Lisp procedures that can serve as the basis for an expert speech recognition system.

We are also using the Lisp machine to run the P3 cognitive simulation language developed by David Zipser and Dan Rabin at UCSD. P3 provides an interactive simulation environment with a graphics interface to model parallel distributed processing systems. We are now using P3 to investigate a wide range of cognitive and perceptual models. First, P3 is currently being used to develop and explore models of auditory word recognition (i.e., the processes by which the pattern of a spoken word is identified). P3 has been used to implement a version of MACS -- a model of the Cohort Theory of auditory word recognition. P3 has also been used to develop alternative models of extent theories of auditory word recognition such as the Logogen model. Work on these models has already provided significant insights into some of the theoretical limitations of these models that would not have been intuitively apparent without a simulation. In addition, P3 is being used to implement the Phonetic Refinement Theory of auditory word recognition that we have developed recently.

A second major use for P3 will be to model the processes used to access lexical knowledge during sentence comprehension. In this project, the goal is to develop a simulation that can, through a set of relatively simple processes, account for a wide range of published data on priming effects and context effects in word perception. Other models of lexical access will be developed to explore the structural and semantic relationships bettween words in the lexicon. Another major function for P3 will be to aid in the development of models of the early perceptual coding processes (i.e., the "front end") used to

process and recognize the acoustic-phonetic structure of fluent speech. For this project, the Spire system will be used to initially code and extract acoustic properties of speech. These features will then be input to a P3 model of acoustic-phonetic processing.

In conclusion then, the DoD-University Instrumentation Grant has provided us with the funds to develop a powerful multi-user digital speech processing system. This new instrumentation has allowed us to investigate a wide range of new and important questions surrounding the perception and production of spoken language. In particular, our work on the effects of noise, alchohol, acceleration and stress on speech production would have been impossible without this facility. In addition, we have been able to develop and test new hypotheses about the structure of words in the lexicon and how this information is used in auditory word recognition. Finally, the Lisp machine provides us with the most sophisticated acoustic-phonetic analysis tools available today, making it possible to investigate the physical and structural properties of speech with greater precision and efficiency than was previously possible.

END

FILMED

3-86

DTIC